

Optimal Explicit Binomial Confidence Interval with Guaranteed Coverage Probability *

Xinjia Chen

Submitted in April, 2008

Abstract

In this paper, we develop an approach for optimizing the explicit binomial confidence interval recently derived by Chen et al. The optimization reduces conservativeness while guaranteeing prescribed coverage probability.

1 Explicit Formula of Chen et al.

Let X be a Bernoulli random variable defined in probability space $(\Omega, \mathcal{F}, \Pr)$ with distribution $\Pr\{X = 1\} = 1 - \Pr\{X = 0\} = p \in (0, 1)$. It is a frequent problem to construct a confidence interval for p based on n i.i.d. random samples X_1, \dots, X_n of X .

Recently, Chen et al. have proposed an explicit confidence interval in [2] with lower confidence limit

$$L_{n,\delta} = \frac{K}{n} + \frac{3}{4} \frac{1 - \frac{2K}{n} - \sqrt{1 + \frac{9}{2 \ln \frac{2}{\delta}} K(1 - \frac{K}{n})}}{1 + \frac{9n}{8 \ln \frac{2}{\delta}}} \quad (1)$$

and upper confidence limit

$$U_{n,\delta} = \frac{K}{n} + \frac{3}{4} \frac{1 - \frac{2K}{n} + \sqrt{1 + \frac{9}{2 \ln \frac{2}{\delta}} K(1 - \frac{K}{n})}}{1 + \frac{9n}{8 \ln \frac{2}{\delta}}} \quad (2)$$

where $K = \sum_{i=1}^n X_i$. Such confidence interval guarantees that the coverage probability $\Pr\{L_{n,\delta} < p < U_{n,\delta} \mid p\}$ is greater than $1 - \delta$ for any $p \in (0, 1)$.

Clearly, the explicit binomial confidence interval is conservative and it is desirable to optimize the confidence interval by tuning the parameter δ . This is objective of the next section.

*The author is currently with Department of Electrical Engineering, Louisiana State University at Baton Rouge, LA 70803, USA, and Department of Electrical Engineering, Southern University and A&M College, Baton Rouge, LA 70813, USA; Email: chenxinjia@gmail.com

2 Optimization of Explicit Binomial Confidence Interval

As will be seen in Section 3, it can be shown that

Theorem 1 *For any fixed n and $p \in (0, 1)$, the coverage probability of confidence interval $[L_{n,\delta}, U_{n,\delta}]$ decreases as δ increases.*

Hence, it is possible to find $\delta > \alpha$ such that

$$\Pr\{L_{n,\alpha} < p < U_{n,\alpha} \mid p\} > 1 - \alpha, \quad \forall p \in (0, 1)$$

for $\alpha \in (0, 1)$. To reduce conservatism of the confidence interval, we consider the following optimization problem:

For a given $\alpha \in (0, 1)$, maximize δ subject to the constraint that

$$\inf_{p \in (0, 1)} \Pr\{L_{n,\delta} \leq p \leq U_{n,\delta} \mid p\} \geq 1 - \alpha.$$

A similar problem is to maximize δ subject to the constraint that

$$\inf_{p \in (0, 1)} \Pr\{L_{n,\delta} < p < U_{n,\delta} \mid p\} \geq 1 - \alpha.$$

As a result of Theorem 1, the maximum δ can be obtained from $(\alpha, 1)$ by a bisection search. In this regard, it is essential to efficiently evaluate $\inf_{p \in (0, 1)} \Pr\{L_{n,\delta} \leq p \leq U_{n,\delta} \mid p\}$ and $\inf_{p \in (0, 1)} \Pr\{L_{n,\delta} < p < U_{n,\delta} \mid p\}$. This is accomplished by the following theorem derived from the theory of random intervals established in [3].

Theorem 2 *Let $\delta \in (0, 1)$. Define*

$$T^-(p) = np + \frac{1 - 2p - \sqrt{1 + \frac{18np(1-p)}{\ln(2/\delta)}}}{\frac{2}{3n} + \frac{3}{\ln(2/\delta)}}, \quad T^+(p) = np + \frac{1 - 2p + \sqrt{1 + \frac{18np(1-p)}{\ln(2/\delta)}}}{\frac{2}{3n} + \frac{3}{\ln(2/\delta)}}$$

for $p \in (0, 1)$ and

$$\mathcal{L}(k) = \frac{k}{n} + \frac{3}{4} \frac{1 - \frac{2k}{n} - \sqrt{1 + \frac{9}{2\ln \frac{2}{\delta}} k(1 - \frac{k}{n})}}{1 + \frac{9n}{8\ln \frac{2}{\delta}}}, \quad \mathcal{U}(k) = \frac{k}{n} + \frac{3}{4} \frac{1 - \frac{2k}{n} + \sqrt{1 + \frac{9}{2\ln \frac{2}{\delta}} k(1 - \frac{k}{n})}}{1 + \frac{9n}{8\ln \frac{2}{\delta}}}$$

for $k = 0, 1, \dots, n$. Define

$$C_l(k) = \Pr\{\lceil T^-(\mathcal{L}(k)) \rceil \leq K \leq k-1 \mid \mathcal{L}(k)\}, \quad C'_l(k) = \Pr\{\lfloor T^-(\mathcal{L}(k)) \rfloor + 1 \leq K \leq k-1 \mid \mathcal{L}(k)\}$$

for $k \in \{0, 1, \dots, n\}$ such that $0 < \mathcal{L}(k) < 1$. Define

$$C_u(k) = \Pr\{k+1 \leq K \leq \lfloor T^+(\mathcal{U}(k)) \rfloor \mid \mathcal{U}(k)\}, \quad C'_u(k) = \Pr\{k+1 \leq K \leq \lceil T^+(\mathcal{U}(k)) \rceil - 1 \mid \mathcal{U}(k)\}$$

for $k \in \{0, 1, \dots, n\}$ such that $0 < \mathcal{U}(k) < 1$.

Then, the following statements hold true:

(I): $\inf_{p \in (0,1)} \Pr\{L_{n,\delta} \leq p \leq U_{n,\delta} \mid p\}$ equals to the minimum of

$$\{C_l(k) : 0 \leq k \leq n; 0 < \mathcal{L}(k) < 1\} \cup \{C_u(k) : 0 \leq k \leq n; 0 < \mathcal{U}(k) < 1\}.$$

(II): $\inf_{p \in (0,1)} \Pr\{L_{n,\delta} < p < U_{n,\delta} \mid p\}$ equals to the minimum of

$$\{C'_l(k) : 0 \leq k \leq n; 0 < \mathcal{L}(k) < 1\} \cup \{C'_u(k) : 0 \leq k \leq n; 0 < \mathcal{U}(k) < 1\}.$$

The proof of Theorem 2 is provided in Section 4.

3 Proof of Theorem 1

For simplicity of notations, define $\lambda = \frac{9n}{8 \ln \frac{2}{\delta}}$ and $z = \frac{k}{n}$. Then, for $K = k$, the upper and lower confidence limits are $U_{n,\delta} = U(z)$ and $L_{n,\delta} = L(z)$ respectively, where

$$U(z) = z + \frac{3}{4} \frac{1 - 2z + \sqrt{1 + 4\lambda z(1 - z)}}{1 + \lambda}, \quad L(z) = z + \frac{3}{4} \frac{1 - 2z - \sqrt{1 + 4\lambda z(1 - z)}}{1 + \lambda}.$$

Since $(1 - 2p)^2 \leq 1 + 4\lambda p(1 - p)$ for $p \in (0, 1)$ and $\lambda > 0$, we have $L(z) \leq z$ and $U(z) \geq z$. Hence, to show Theorem 1, it suffices to show that both $U(z) - z$ and $z - L(z)$ decrease as δ increases for any $z \in [0, 1]$. We shall first show that $U(z) - z$ decreases as δ increases for any fixed $z \in [0, 1]$. For this purpose, we can define

$$y = \frac{4[U(z) - z]}{3}$$

and show that $\frac{\partial y}{\partial \lambda} < 0$. To this end, we can use the definition of y to obtain the following equation $[(1 + \lambda)y - (1 - 2z)]^2 = 1 + 4\lambda z(1 - z)$. Differentiating both sides of this equation with respect to λ yields

$$2[(1 + \lambda)y - (1 - 2z)] \left[(1 + \lambda) \frac{\partial y}{\partial \lambda} + y \right] = 4z(1 - z),$$

from which we have

$$(1 + \lambda) \frac{\partial y}{\partial \lambda} = \frac{2z(1 - z)}{(1 + \lambda)y - (1 - 2z)} - y.$$

Clearly, to show $\frac{\partial y}{\partial \lambda} < 0$, it suffices to show that the right-hand side of the above equation is negative for any $z \in [0, 1]$ and $\lambda > 0$. That is, to show

$$\frac{2z(1 - z)}{\sqrt{1 + 4\lambda z(1 - z)}} < y,$$

or equivalently,

$$(1 + \lambda)2z(1 - z) < (1 - 2z)\sqrt{1 + 4\lambda z(1 - z)} + 1 + 4\lambda z(1 - z),$$

which can be written as $2z(1-z) < w(\lambda)$, where

$$w(\lambda) = (1-2z)\sqrt{1+4\lambda z(1-z)} + 1 + 2\lambda z(1-z).$$

Note that

$$\frac{\partial w(\lambda)}{\partial \lambda} = 2z(1-z) \left[\frac{1-2z}{\sqrt{1+4\lambda z(1-z)}} + 1 \right] > 0$$

as a result of

$$\frac{1-2z}{\sqrt{1+4\lambda z(1-z)}} > -\frac{1}{\sqrt{1+4\lambda z(1-z)}} > -1.$$

Hence, $w(\lambda) > w(0) = 2(1-z)$ for any $\lambda > 0$. This shows that $2z(1-z) < w(\lambda)$ for any $z \in [0, 1]$ and $\lambda > 0$. Consequently, we have established $\frac{\partial y}{\partial \lambda} < 0$, which implies that $U(z) - z$ decreases as δ increases for any fixed $z \in [0, 1]$. Observing that $L(z) = 1 - U(1-z)$ for any $z \in [0, 1]$, we have

$$z - L(z) = z - [1 - U(1-z)] = U(1-z) - (1-z).$$

Therefore, it must be true that $z - L(z)$ decreases as δ increases for fixed any $z \in [0, 1]$. So, the proof of Theorem 1 is completed.

4 Proof of Theorem 2

For simplicity of notations, we define λ , z , $L(z)$ and $U(z)$ as in the proof of Theorem 1. Note that $L(1) = 1 - \frac{3}{2(1+\lambda)} < 1$ and the derivative of $L(z)$ with respect to z is

$$\begin{aligned} L'(z) &= 1 + \frac{3}{4(1+\lambda)} \left[-2 - \frac{1}{2} \frac{4\lambda(1-2z)}{\sqrt{1+4\lambda z(1-z)}} \right] \\ &= 1 - \frac{3}{2(1+\lambda)} - \frac{3}{2(1+\lambda)} \frac{\lambda(1-2z)}{\sqrt{1+4\lambda z(1-z)}} \\ &= \frac{1}{2(1+\lambda)} \left[2\lambda - 1 - \frac{3\lambda(1-2z)}{\sqrt{1+4\lambda z(1-z)}} \right] \end{aligned}$$

which is positive if and only if $(2\lambda - 1)\sqrt{1+4\lambda z(1-z)} > 3\lambda(1-2z)$.

To complete the proof of Theorem 2, we need some preliminary results.

Lemma 1 For any $n \geq 1$ and $\delta \in (0, 1)$,

$$\frac{(2\lambda - 1)^2(1 + \lambda)}{36\lambda^2 + 4\lambda(2\lambda - 1)^2} \geq \frac{1}{4} \quad (3)$$

if and only if $\lambda \leq \frac{1}{5}$.

Proof. Note that (3) is equivalent to $(2\lambda - 1)^2(1 + \lambda) \geq 9\lambda^2 + \lambda(2\lambda - 1)^2$, which can be simplified as $(5\lambda - 1)(\lambda + 1) \leq 0$. Since $\delta \in (0, 1)$ and $n \geq 1$, we have $\lambda > 0$. Hence, the inequality (3) holds if and only if $\lambda \leq \frac{1}{5}$. □

Lemma 2 $L(z)$ is monotonically increasing with respect to $z \in [0, 1]$ such that $L(z) > 0$. Similarly, $U(z)$ is monotonically increasing with respect to $z \in [0, 1]$ such that $U(z) < 1$.

Proof. We shall first show that $L(z)$ is monotonically increasing with respect to $z \in [0, 1]$ such that $L(z) > 0$. It suffices to consider four cases:

- Case (i): $\lambda \geq \frac{1}{2}$ and $0 < z \leq \frac{1}{2}$;
- Case (ii): $\lambda \geq \frac{1}{2}$ and $1 > z > \frac{1}{2}$;
- Case (iii): $\lambda < \frac{1}{2}$ and $0 < z \leq \frac{1}{2}$;
- Case (iv): $\lambda < \frac{1}{2}$ and $1 > z > \frac{1}{2}$.

In Case (i), $L(z)$ increases if and only if $(2\lambda - 1)^2[1 + 4\lambda z(1 - z)] > 9\lambda^2(1 - 2z)^2$, or equivalently,

$$\left(z - \frac{1}{2}\right)^2 < \frac{(2\lambda - 1)^2(1 + \lambda)}{36\lambda^2 + 4\lambda(2\lambda - 1)^2}.$$

Define

$$z^* = \frac{1}{2} - \sqrt{\frac{(2\lambda - 1)^2(1 + \lambda)}{36\lambda^2 + 4\lambda(2\lambda - 1)^2}}.$$

By Lemma 1, we have $z^* > 0$. It follows that $L(z)$ is monotonically decreasing with respect to $z \in (0, z^*)$ and monotonically increasing with respect to $z \in (z^*, \frac{1}{2})$. This implies that $L(z)$ achieves its minimum at z^* and $L(z) < L(0) = 0$ for any $z \in (0, z^*)$. Therefore, we have shown that $L(z)$ is monotonically increasing with respect to $z \in (0, 1)$ such that $L(z) \geq 0$ and that the conditions of Case (i) hold true.

In Case (ii), $L(z)$ increases for $z \in (\frac{1}{2}, 1)$.

In Case (iii), $L(z)$ decreases for $z \in (0, \frac{1}{2}]$. It can be seen that $L(z) < L(0) = 0$ for any $z \in (0, \frac{1}{2}]$.

In Case (iv), $L(z)$ increases if and only if $(2\lambda - 1)\sqrt{1 + 4\lambda z(1 - z)} > 3\lambda(1 - 2z)$, which can be written as $(1 - 2\lambda)\sqrt{1 + 4\lambda z(1 - z)} < 3\lambda(2z - 1)$ or equivalently,

$$\left(z - \frac{1}{2}\right)^2 > \frac{(2\lambda - 1)^2(1 + \lambda)}{36\lambda^2 + 4\lambda(2\lambda - 1)^2}$$

Define

$$z^* = \frac{1}{2} + \sqrt{\frac{(2\lambda - 1)^2(1 + \lambda)}{36\lambda^2 + 4\lambda(2\lambda - 1)^2}}.$$

If $\frac{1}{2} > \lambda > \frac{1}{5}$, by Lemma 1, we have $z^* < 1$. Hence, $L(z)$ increases for $z \in (z^*, 1)$ and

$$L(z) < L(1) = z + \frac{3}{4} \frac{1 - 2z - \sqrt{1 + 4\lambda z(1 - z)}}{1 + \lambda} = \frac{2\lambda - 1}{2(1 + \lambda)} < 0, \quad \forall z \in (z^*, 1).$$

Moreover, $L(z)$ decreases for $z \in (\frac{1}{2}, z^*)$ and

$$L(z) < L\left(\frac{1}{2}\right) = z + \frac{3}{4} \frac{1 - 2z - \sqrt{1 + 4\lambda z(1 - z)}}{1 + \lambda} = \frac{1}{2} - \frac{3}{4\sqrt{1 + \lambda}} < 0, \quad \forall z \in (z^*, 1).$$

If $0 < \lambda \leq \frac{1}{5}$, by Lemma 1, we have $z^* \geq 1$. Hence, $L(z)$ decreases for $z \in (\frac{1}{2}, 1)$ and

$$L(z) < L\left(\frac{1}{2}\right) < 0 \quad \forall z \in \left(\frac{1}{2}, 1\right).$$

Based on the preceding investigation, we can conclude that the lower confidence limit is non-decreasing with respect to $z \in (0, 1)$ such that $L(z) \geq 0$. Recalling that $L(1) < 1$, we have that $L(z) < 1$ for any $z \in (0, 1)$.

Since $U(z) = 1 - L(1 - z) > 0$ for any $z \in (0, 1)$, we have that the upper confidence limit $U(z)$ is also non-decreasing with respect to $z \in (0, 1)$ such that $U(z) \leq 1$.

□

Now we consider the minimum coverage probability. By the definitions of $L_{n,\delta}$, $U_{n,\delta}$ and $\mathcal{L}(k)$, $\mathcal{U}(k)$, we have

$$\Pr\{L_{n,\delta} \leq p < U_{n,\delta} \mid \mathcal{U}(k)\} = \Pr\{k < K \leq T^+(p) \mid \mathcal{U}(k)\}, \quad 0 < \mathcal{U}(k) < 1$$

$$\Pr\{L_{n,\delta} < p \leq U_{n,\delta} \mid \mathcal{L}(k)\} = \Pr\{T^-(p) \leq K < k \mid \mathcal{L}(k)\}, \quad 0 < \mathcal{L}(k) < 1$$

$$\Pr\{L_{n,\delta} < p < U_{n,\delta} \mid \mathcal{U}(k)\} = \Pr\{k < K < T^+(p) \mid \mathcal{U}(k)\}, \quad 0 < \mathcal{U}(k) < 1$$

$$\Pr\{L_{n,\delta} < p < U_{n,\delta} \mid \mathcal{L}(k)\} = \Pr\{T^-(p) < K < k \mid \mathcal{L}(k)\}, \quad 0 < \mathcal{L}(k) < 1.$$

Since both $L(z)$ and $U(z)$ are monotone, the proof of Theorem 2 can be completed by making use of the above results and applying the theory of coverage probability of random intervals established by Chen in [3].

References

- [1] Brown, L. D., Cai, T. and DasGupta, A., “Interval estimation for a binomial proportion and asymptotic expansions,” *The Annals of Statistics*, vol. 30, pp. 160-201, 2002.
- [2] Chen X., Zhou K. and Aravena J., “Explicit formula for constructing binomial confidence interval with guaranteed coverage probability,” *Communications in Statistics – Theory and methods*, vol. 37, pp. 1173-1180, 2008.
- [3] Chen X., “Coverage Probability of Random Intervals,” arXiv:0707.2814, July 2007.